DISK ARRAY DEVICE

Patent Number: JP7200190
Publication date: 1995-08-04
Inventor(s): TAKAI RIICHI
Applicant(s):: NEC CORP
Requested Patent: JP7200190
Application Number: JP19930355494 19931229
Priority Number(s):
IPC Classification: G06F3/06 ; G06F12/08
EC Classification:
Equivalents: JP2570614B2

Abstract

PURPOSE:To prevent the read performance from decreasing if one disk gets out of order in a disk array which stores data decentralized and stored on plural disks and their parity information.
CONSTITUTION:A cache memory 300 which holds data of respective disk drives 401-40N normally functions as a virtual disk which holds the contents of a faulty disk drive if the fault of one of the disk drives 401-40N is detected.

(19)日本国特許庁（ＪＰ）　(12)公開特許公報（Ａ）　(11)特許出願公開番号

特開平7－200190

(43)公開日　平成7年（1995）8月4日

(51)Int.Cl.[6]		識別記号	庁内整理番号	ＦＩ	技術表示箇所
Ｇ０６Ｆ	3/06	５４０			
		３０５　Ｃ			
	12/08	３２０	7608－5Ｂ		

審査請求　有　請求項の数２　書面（全６頁）

(21)出願番号　特願平5－355494

(22)出願日　平成5年（1993）12月29日

(71)出願人　000004237
日本電気株式会社
東京都港区芝五丁目7番1号

(72)発明者　高井　利一
東京都港区芝五丁目7番1号　日本電気株式会社内

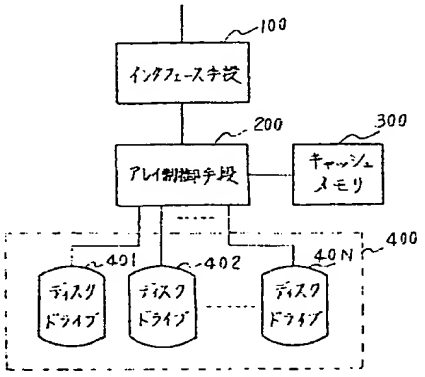
(74)代理人　弁理士　京本　直樹　（外２名）

(54)【発明の名称】デイスクアレイ装置

(57)【要約】

【目的】複数のディスクに分散して格納したデータ及びそれらのパリティ情報を保存するディスクアレイにおいて、１台のディスクが故障した際の読み出し性能低下を防止する。

【構成】正常時にはディスクドライブ４０１～４０Ｎのそれぞれのデータを保持するキャッシュメモリ３００は、ディスクドライブ４０１～４０Ｎのいずれかにおいて故障が検出されると、その故障したディスクドライブの内容を保持する仮想ディスクとしても機能する。



1

【特許請求の範囲】

【請求項１】複数のディスクドライブを有するディスクドライブ群と、

これらディスクドライブ群への書込みデータに関するエラー修復のための情報としてエラー修復情報を生成し、前記ディスクドライブ群からの読出しデータに関してエラーの発生をチェックしてエラー修復をするアレイ制御手段と、

前記複数のディスクドライブの１つが故障した場合にはこの故障したディスクドライブが格納すべきデータとして前記アレイ制御手段により修復されたデータを格納するキャッシュメモリ手段とを含むことを特徴とするディスクアレイ装置。

【請求項２】前記キャッシュメモリ手段は前記故障したディスク以外のディスクが格納するデータをも格納することを特徴とする請求項１に記載のディスクアレイ装置。

【発明の詳細な説明】

【０００１】

【産業上の利用分野】本発明は、情報処理装置に於いて用いる磁気ディスクを用いた冗長性を持つディスクアレイに関し、特に故障したディスクが存在する場合のデータアクセス性能の向上に関する。

【従来の技術】従来情報処理装置において用いられる外部記憶装置はシステム側の必要によりより大容量の記憶装置を用いる方向に進んでいたが、この大容量の外部記憶装置を使用するシステム構成では１つの記憶装置の故障により大量のデータが一度に喪失してしまう。また、同一外部記憶装置の別領域に記憶されているデータへの並列アクセスが出来ないため、システム全体としてのスループットが低下してしまう。上記のような単一の大容量外部記憶装置を用いた場合の弊害をなくすために考え出された方式がカリフォルニア大学バークレー校コンピュータ科学学科により提唱された“低価格ディスクによる冗長アレイ（Ｒｅｄｕｎｄａｎｔ　Ａｒｒａｙｓ　ｏｆ　Ｉｎｅｘｐｅｎｓｉｖｅ　Ｄｉｓｋ、以下ＲＡＩＤという）”であり、例えば、「Ａ　Ｃａｓｅ　ｆｏｒ　Ｒｅｄｕｎｄａｎｔ　Ａｒｒａｙｓｏｆ　Ｉｎｅｘｐｅｎｓｉｖｅ　Ｄｉｓｋｓ　（ＲＡＩＤ），カリフォルニア大学バークレー校コンピュータ科学学科（ＥＥＳＣ），レポート番号ＵＣＢ／ＣＳＤ／８７／３９１」に記載されている。このＲＡＩＤについてはＲＡＩＤ１よりＲＡＩＤ５までの５種類の実現方法が提唱されているが、このうちＲＡＩＤ４及びＲＡＩＤ５として規定されている方式ではＲＡＩＤアレイに対するデータをブロック単位で分割し、その分割したデータをアレイを構成する各ドライブに分散させて記憶し、そのパリティデータを別ドライブに記憶している。このＲＡＩＤ４、５に対し書き込みを行なう場合、該当データとそれに対応するパリティデータを一度読みだしてから再書き込みを行なう必要がある。この為通常の単体ドライブに対するデータ書き込みに比べ余分なディスクアクセスが必要となる。この読み出し及び再書き込みのために必要となるオーバーヘッドをライトペナルティと呼ぶ。このライトペナルティを少なくするため通常はホストからの書き込みデータがアレイを構成している磁気ディスクへ書き込まれる前にキャッシュメモリに一度データを保持し、実際に磁気ディスクへのデータ書き込みが終了する前にアレイとしてホストへ書き込み終了を通知することにより書き込み時の性能低下を防いでいる。

2

【０００２】また、現在実用化されているディスクアレイにおいて１台のドライブが故障した場合、他のドライブのデータより故障したドライブのデータを修復する必要があるためアレイ全体の性能が低下する。これを防ぐためには故障したドライブに記憶されているはずのデータと同等のデータを持つ仮想ドライブを用意する必要がある。

【０００３】

【発明が解決しようとする課題】上述したようにディスクアレイでは自己を構成する磁気ディスク装置が故障した縮退状態に陥っている場合、故障した磁気ディスク装置に記憶されているデータを他の磁気ディスク装置より再構築する必要があるためどうしても通常時に比べ読み出し速度が低下してしまう。

【０００４】本発明はディスクアレイの書き込み性能の向上のために用意されているライトキャッシュメモリを故障した磁気ディスク装置とデータ的に同等な仮想ディスクとして用いることにより縮退時のデータアクセス性能の低下を抑えることを目的とする。

【０００５】

【課題を解決するための手段】上記課題を解決するため、本願発明のディスクアレイ装置では、複数のディスクドライブを有するディスクドライブ群と、これらディスクドライブ群への書込みデータに関するエラー修復のための情報としてエラー修復情報を生成し、前記ディスクドライブ群からの読出しデータに関してエラーの発生をチェックしてエラー修復をするアレイ制御手段と、前記複数のディスクドライブの１つが故障した場合にはこの故障したディスクドライブが格納すべきデータとして前記アレイ制御手段により修復されたデータを格納するキャッシュメモリ手段とを有している。

【０００６】また、前記キャッシュメモリ手段は前記故障したディスク以外のディスクが格納するデータをも格納する。

【０００７】

【実施例】次に本願発明のディスクアレイ装置の一実施例について図面を参照して詳細に説明する。

【０００８】図１を参照すると、本願発明の一実施例であるディスクアレイ装置は、インタフェース手段１００と、アレイ制御手段２００と、キャッシュメモリ３００

と、ディスクドライブ群４００とを含んでいる。

【０００９】インタフェース手段１００は、他の装置からアクセス要求を受け、また、返答を返す。

【００１０】アレイ制御手段２００は、ディスクアレイ装置全体の制御を行なう。すなわち、ディスクドライブ群４００にアクセスすると共に、キャッシュメモリ３００の管理を行なう。

【００１１】キャッシュメモリ３００は、ディスクドライブ群４００の内容を一時的に格納するディスクキャッシュの機能を有し、また、後述するように故障したディスクに相当するデータを格納する。

D40N= D401 xor D402 xor D403 xor … xor D40(N-1)　…式（１）

で表されるものとする。但し、“ｘｏｒ”は排他的論理和を示し、Ｄ４０ｘはそれぞれディスク４０ｘの対応するデータを示す。ディスク４０Ｎがこの様に定義された場合、任意のディスクに保持されているデータは他の（Ｎ－１）台のディスクに保持されているデータの排他的論理和と等しくなるため、１台のディスクが故障した場合でも他の（Ｎ－１）台のディスクのデータを読み出しそのデータの排他的論理和を求めればデータ復旧を行うことが可能となる。通常、ディスクとのデータアクセスはセクタと呼ばれるあらかじめ決められたデータサイズを単位として行われるが、ＲＡＩＤアレイに対しセク

Nnew= (Aold xor Anew) xor Nold　…式（２）

であらわすことができることを利用し、パリティディスク４０Ｎとディスク４０Ａに対する読み出しと書き込みだけで新たなデータの書き込みを終了できる。但し、この処理は単独のディスクに対する書き込みならば１回行なえば済んでいたディスクアクセスが、ディスクアレイでは書き込み処理中に、（１）パリティディスク４０ＮからのＮｏｌｄの読み出し、（２）データディスク４０ＡからのＡｏｌｄの読み出し、（３）パリティディスク４０ＮへのＮｎｅｗの書き出し、（４）データディスク４０ＡへのＡｎｅｗの書き出し、の４回のアクセスを行なう必要が生じるため、１回の書き込み終了時間が遅くなる。以下、これをライトペナルティとよぶ。

【００１４】本願発明のディスクアレイ装置の一実施例

D40X= D401 xor … D40(X-1) xor D40(X+1) … xor D40N　…式（３）

と言う関係が成り立つため、他の（Ｎ－１）台のディスクデータの排他的論理和を計算することで、故障ディスク４０Ｘのデータを復旧することが可能となる。なお、このようにディスクの１台が故障し、他のディスクのデータからデータを計算により算出している状態を縮退状態と呼ぶ。一般に、この縮退状態では通常は４０Ｘのディスクへのアクセスですむデータ読み出しを行う代わりに、４０Ｘを除く（Ｎ－１）台のディスクへのアクセスが必要となるため、正常状態よりもアクセス性能が落ちる。この性能低下を防止するため、本願発明のディスクアレイ装置の一実施例では、縮退状態においては、キャッシュメモリ３００が故障ディスク４０Ｘに相当する内

【００１２】ディスクドライブ群４００は、さらにＮ台のディスクドライブ４０１～４０Ｎを含んでいる。ＲＡＩＤ構成をとるディスクアレイの場合、上記Ｎ台のディスクのうちの１台をパリティデータ保持に使用する。本実施例においてはディスク４０Ｎをパリティデータ保持に使用するものとする。

【００１３】本願発明のディスクアレイ装置の一実施例では、通常のＲＡＩＤ構成を持つディスクアレイ装置と同様に、パリティデータ保持に使用されるディスク４０Ｎに書き込まれるデータＤ４０Ｎは、

タの大きさ以下の書き込みを行う場合には、複数のディスクへの同時アクセスを可能とするために、書き込み時に全てのディスクのデータを書き換えるのではなく、書き込みデータの保持を行なうディスク４０Ａとパリティディスク４０Ｎにのみアクセスを行う必要がある。すなわち、ディスク４０Ａに既に書かれているデータをＡｏｌｄ，これから書き込まれるデータをＡｎｅｗとし、パリティディスク４０Ｎに書かれていたデータをＮｏｌｄとした場合、パリティディスクに新たに書き込む必要のあるデータＮｎｅｗは他のディスクに書き込まれているデータには関係なく、

では、通常のＲＡＩＤ構成を持つディスクアレイ装置と同様に、このライトペナルティを短縮するために、一時的にキャッシュメモリ３００に書き込みデータを保持し、実際にディスク４０１，４０２…４０Ｎにデータが書き込まれていなくともキャッシュメモリ３００にデータが書き込まれた時点で、インタフェース手段１００に対し終了報告を行う方式を採用する。

【００１５】このように構成することにより、任意のディスク４０Ｘが故障した場合（以下、この故障したディスク４０Ｘを故障ディスクという）、通常の単独ディスクではそのディスク４０Ｘに格納されていたデータを復旧することは不可能であるが、ＲＡＩＤ構成を持つアレイでは式（１）より、

容を格納することにより、故障ディスク４０Ｘへのアクセス時に復旧をしないで済むことになる。

【００１６】図２（Ａ）を参照すると、本発明の一実施例のディスクアレイ装置が正常なディスクアレイとして動作している場合のキャッシュメモリ３００のメモリマップにおいて、キャッシュメモリ３００は各ディスク５０１，５０２…５０Ｎに書き込まれるべきデータを一時的に保持している。インタフェース手段１００より転送された書き込みデータはディスク５０１，５０２…５０Ｎへの書き込み前に各々のディスクに対応させたディスクキャッシュ３０１，３０２…３０Ｎに一度書き込まれ、この段階でインタフェース手段１００を介して、上

位装置に対し終了報告を行うことにより物理動作が必要なディスク４０１，４０２…４０Ｎへデータを書き込む場合に比べ、システム全体の書き込み性能の向上を図れる。

【００１７】図２（Ｂ）を参照すると、縮退状態のキャッシュメモリ３００のメモリマップにおいて、キャッシュメモリ３００は書き込みデータ用のディスクキャッシュ３０１，３０２…３０Ｎとしてではなく、故障したディスク４０Ｘと等価な仮想ディスク７００として位置付けられ、故障ディスク４０Ｘに対するデータ読み出し時に再構成されたデータを仮想ディスク７００に記憶させる。つまり、正常時はアレイに書き込まれたデータを保存し次に書き込まれたデータを読み出す場合にメモリ上のデータを使用することにより１回のディスクアクセスを無くしているが、縮退時には故障しているディスク４０Ｘのデータを式（３）により再構成した結果を保存するために使用することにより故障ディスクデータの読み出し時にに必要な（Ｎ－１）台のディスクへのアクセスを減らすことを可能とする。この様に一度読み出し処理が行われた事により再構成された故障ディスク４０Ｘと等価なデータを仮想ディスク７００に蓄積し、同一領域を再度読み出す場合にはこの蓄積されているデータを読み出しデータとして使用することにより縮退状態における読み出し性能の低下を防ぐことが可能となる。この場合、キャッシュメモリ３００の容量はディスク４０１，４０２…４０Ｎと同等以上であることが望ましいが、ディスクより容量が少ない場合でも、通常のキャッシュメモリと同様の機能が働き、使用頻度の高いデータを優先して保持することになるため、仮想ディスク７００に記憶されているデータのヒット率は一定のレベルに保つことが可能である。

【００１８】次に本願発明の一実施例のディスクアレイ装置の動作について説明する。

【００１９】図３を参照すると、本願発明の一実施例のディスクアレイ装置の縮退時の読出し動作を表す処理の流れ図である。上位装置からのアクセス要求はインタフェース手段１００を介してアレイ制御手段２００に渡され、このアレイ制御手段２００で解読される。このアクセス要求が読み出し要求であれば、アレイ制御手段２００は図３の流れ図に従って、故障ディスク４０Ｘに対するアクセスか否かが判断される（ステップ２０１）。故障ディスク４０Ｘ以外のディスクドライブに対する読み出し要求であれば、アレイ制御手段２００は該当するディスクドライブから読み出しを行なう（ステップ２０２）。

【００２０】故障ディスク４０Ｘに対する読み出し要求であれば（ステップ２０１）、当該データがキャッシュメモリ３００に格納されているか否かが調べられる（ステップ２０３）。当該データがキャッシュメモリ３００に格納されていれば、すなわちキャッシュヒットであれば、当該データがキャッシュメモリ３００から読み出される（ステップ２０４）。

【００２１】故障ディスク４０Ｘに対する読み出し要求であって、しかも当該データがキャッシュメモリ３００に格納されていなければ（ステップ２０３）、アレイ制御手段２００はディスクドライブ群４００から当該アクセス要求に対応するパリティ、すなわち式（１）が成立するＤ４０Ｎ、をパリティディスク４０Ｎから読み出す（ステップ２０５）。

【００２２】続いて、アレイ制御手段２００は当該アクセス要求に係るデータに対応するデータ群、すなわち式（１）が成立するＤ４０１～Ｄ４０（Ｎ－１）、をディスクドライブ群４００から読み出す（ステップ２０６）。但し、故障ディスク４０Ｘからはデータを読み出さない。

【００２３】そして、上記パリティＤ４０Ｎとデータ群Ｄ４０１～Ｄ４０（Ｎ－１）とから、式（３）によって、故障ディスク４０Ｘに対応するデータＤ４０Ｘを復旧する（ステップ２０７）。このようにして復旧されたデータＤ４０Ｘは、キャッシュメモリ３００に書き込まれる（ステップ２０８）。

【００２４】図４を参照すると、本願発明の一実施例のディスクアレイ装置の縮退時の書込み動作を表す処理の流れ図である。上位装置からのアクセス要求が書込み要求であれば、アレイ制御手段２００は図４の流れ図に従って、ディスクドライブ群４００から当該アクセス要求に対応するパリティ、すなわち式（１）が成立するＤ４０Ｎ、をパリティディスク４０Ｎから読み出す（ステップ２１１）。

【００２５】続いて、アレイ制御手段２００は当該アクセス要求に係るデータに対応するデータ群、すなわち式（１）が成立するＤ４０１～Ｄ４０（Ｎ－１）、をディスクドライブ群４００から読み出す（ステップ２１２）。但し、故障ディスク４０Ｘからはデータを読み出さない。

【００２６】そして、アレイ制御手段２００は書込みデータを新たに含めたデータ群Ｄ４０１～Ｄ４０（Ｎ－１）から式（１）によってパリティＤ４０Ｎを生成して、このパリティＤ４０Ｎをパリティディスク４０Ｎに書き込む（ステップ２１３）。

【００２７】その後、書込みデータが故障ディスク４０Ｘに対するものであればこの書込みデータはキャッシュメモリ３００に書き込まれ（ステップ２１６）、故障ディスク４０Ｘ以外のディスクドライブに対するものであれば当該ディスクドライブに書き込まれる（ステップ２１５）。

【００２８】図５及び図６を参照すると、図５はキャッシュメモリ３００をディスクキャッシュとして用いた場合のデータ転送ルートを示す図であり、図６は縮退時にキャッシュメモリ３００を仮想ディスク７００として使

用した場合のデータ転送ルートを示す図である。

【００２９】なお、縮退時にキャッシュメモリ３００を仮想ディスク７００として用いた場合、前に示したライトペナルティを避ける方法が無くなるためにより書き込み性能は低下する。この性能低下は、図２（Ｃ）にメモリマップを示したように縮退時において常にキャッシュメモリ３００の全容量を仮想ディスク７００として確保するのではなく、故障ディスクと等価なディスクとして再構成されたデータを保存していない部分をディスクキャッシュとして用いることにより防ぐことが可能となる。つまり、キャッシュメモリ３００上にライトペナルティを防ぐキャッシュとして利用する領域３０１～３０Ｎと仮想ディスク７００として利用する領域を双方用意することにより、仮想ディスク領域７００で縮退時の読み出し性能の低下を、キャッシュ領域３０１～３０Ｎでライトペナルティの防止を行なうことが可能となる。

【００３０】

【発明の効果】以上説明したように、本発明はディスクアレイにおけるライトペナルティの補完用であるキャッシュメモリを、縮退した磁気ディスク装置の仮想ディスクとして用いることにより、縮退したディスクアレイからのデータ読み出し時の性能低下を抑えることを可能とする効果を有する。

【図面の簡単な説明】

【図１】本発明のディスクアレイ装置の一実施例を表すブロック図である。

【図２】本発明の一実施例のキャッシュメモリが格納するデータのメモリマップを表す図である。

【図３】本発明のディスクアレイ装置の一実施例の読出し動作の処理の流れを表す図である。

【図４】本発明のディスクアレイ装置の一実施例の書込み動作の処理の流れを表す図である。

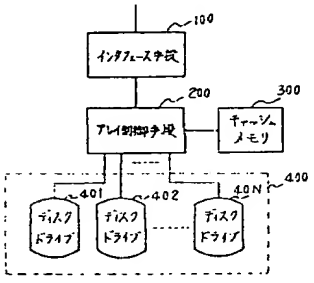
【図５】本発明のディスクアレイ装置の一実施例における正常時のデータ転送ルートを表す図である。

【図６】本発明のディスクアレイ装置の一実施例における故障発生時のデータ転送ルートを表す図である。

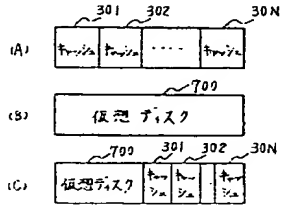
【符号の説明】

１００　インタフェース手段
２００　アレイ制御手段
３００　キャッシュメモリ
３０１～３０Ｎ　ディスクキャッシュ領域
４００　ディスクドライブ群
４０１～４０Ｎ　ディスクドライブ
７００　仮想メモリ領域

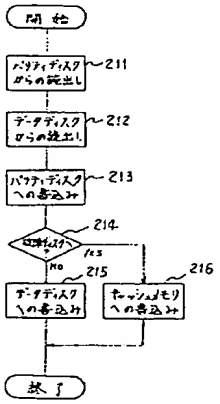
【図１】



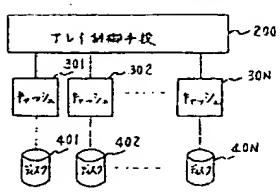
【図２】



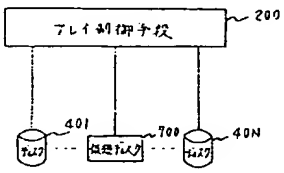
【図４】



【図５】



【図６】



【図3】